<Bioinformatics Center>Mathematical Bioinformatics

Bioinformatics Center
– Mathematical Bioinformatics –

http://www.bic.kyoto-u.ac.jp/takutsu/index.html

Prof
AKUTSU, Tatsuya
(D Eng)

Assist Prof
HAYASHIDA, Morihiro
(D Inf)

Assist Prof
TAMURA, Takeyuki
(D Inf)

PD (JSPS)
ZHAO, Yang
(D Inf)

Students

NAKAJIMA, Natsu (D3)
LU, Wei (D3)
UECHI, Risa (D3)
RUAN, Peiying (D3)
HASEGAWA, Takanori (D3)
MORI, Tomoya (D3)
JIRA, Jindalertudomdee (D2)
BAO, Yu (D1)
HWANG, Jaewook (M2)
CAO, Yue (M1)
NGOUV, Hayliang (M1)

Guest Scholar

CHENG, Xiaoqing The University of Hong Kong, China P.R., 20 May-15 August

Guest Res Assoc

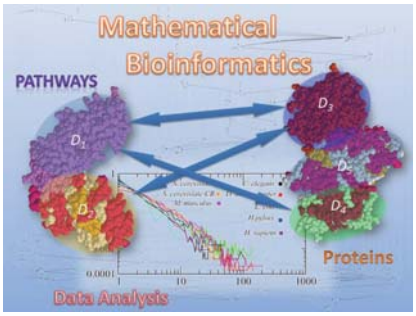
LIU, Liwei (Ph D) Dalian University of Thechnology, China P.R., 1 April-31 March

Scope of Research

Due to rapid progress of the genome projects, whole genome sequences of organisms ranging from bacteria to human have become available. In order to understand the meaning behind the genetic code, we have been developing algorithms and software tools for analyzing biological data based on advanced information technologies such as theory of algorithms, artificial intelligence, and machine learning. We are recently studying the following topics: systems biology, scalefree networks, protein structure prediction, inference of biological networks, chemo-informatics, discrete and stochastic methods for bioinformatics.

KEYWORDS

Scale-free Networks
Boolean Networks
Chemical Graphs
Grammar-based Compression
Protein Complexes



Selected Publications

Akutsu, T.; Tamura, T.; Fukagawa, D.; Takasu, A., Efficient Exponential-Time Algorithms for Edit Distance between Unordered Trees, *Journal of Discrete Algorithms*, **25,** 79-93 (2014).
Hayashida, M.; Ruan, P.; Akutsu, T., Proteome Compression via Protein Domain Compositions, *Methods*, **67,** 380-385 (2014).
Lu, W.; Tamura, T.; Song, J.; Akutsu, T., Integer Programming-based Method for Designing Synthetic Metabolic Networks by Minimum Reaction Insertion in a Boolean Model, *PLoS ONE*, **9,** e92637 (2014).
Suzuki, M.; Nagamochi, H.; Akutsu, T., Efficient Enumeration of Monocyclic Chemical Graphs with Given Path Frequencies, *Journal of Cheminformatics*, **6,** 31 (2014).
Wang, M.; Zhao, X.-M.; Tan, H.; Akutsu, T.; Whisstock J. C.; Song, J., Cascleave 2.0, A New Approach for Predicting Caspase and Granzyme Cleavage Targets, *Bioinformatics*, **30,** 71-80 (2014).

Proteome Compression via Protein Domain Compositions

We focus on the entropy that the individual contains, and analyze domain compositions of proteins through compression of whole proteins in an organism. We suppose that a protein is a multiset of domains. Since gene duplication and fusion have occurred through evolutionary processes, the same domains and the same compositions of domains appear in multiple proteins, which enables us to compress a proteome by using references to proteins for duplicated and fused proteins. Such a network with references to at most two proteins is modeled as a directed hypergraph.

We propose a heuristic approach by combining the Edmonds algorithm with an integer linear programming, and apply our procedure to fourteen proteomes of *D. discoideum, E. coli, S. cerevisiae, S. pombe, C. elegans, D. melanogaster, A. thaliana, O. sativa, D. rerio, X. laevis, G. gallus, M. musculus, P. troglodytes, and H. sapiens*.

The compressed size using both of duplication and fusion was smaller than that using only duplication, which suggests the importance of fusion events in evolution of a proteome. In addition, we observed the difference of gene duplication rates between organisms. It is considered that gene duplication in *M. musculus and H. sapiens* tends to occur more frequently than other organisms examined in this study. Furthermore, we observed the phenomenon in several organisms that a fused gene was used in another gene fusion event again. For correlation between the compression ratio of each proteome and the phylogenetic tree, further analysis is needed. The proteome compression using domain compositions in this study can be applied to compression of protein amino acid sequences and DNA base sequences, and the compression ratio may be improved by making use of sequences included in domains as reference.

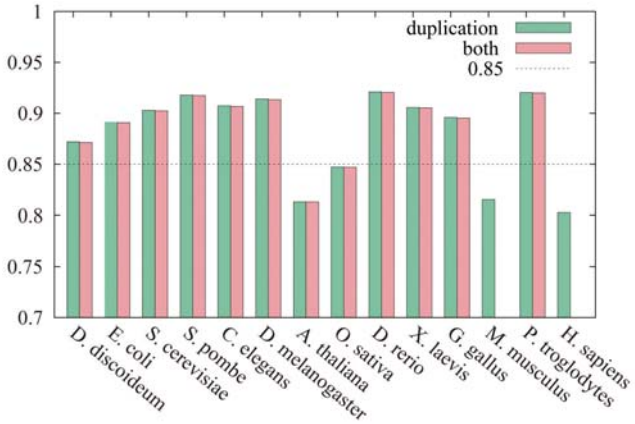


Figure 1. Results on the compression ratio by duplication rules and by both duplication and fusion rules.

Integer Programming-Based Method for Designing Synthetic Metabolic Networks by Minimum Reaction Insertion in a Boolean Model

In this study, we consider the Minimum Reaction Insertion (MRI) problem for finding the minimum number of additional reactions from a reference metabolic network to a host metabolic network so that a target compound becomes producible in the revised host metabolic network in a Boolean model. Although a similar problem for larger networks is solvable in a flux balance analysis (FBA)-based model, the solution of the FBA-based model tends to include more reactions than that of the Boolean model. However, solving MRI using the Boolean model is computationally more expensive than using the FBA-based model since the Boolean model needs more integer variables. Therefore, in this study, to solve MRI for larger networks in the Boolean model, we have developed an efficient Integer Programming formalization method in which the number of integer variables is reduced by the notion of feedback vertex set and minimal valid assignment. As a result of computer experiments conducted using the data of metabolic networks of E. coli and reference networks downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, we have found that the developed method can appropriately solve MRI in the Boolean model and is applicable to large scale-networks for which an exhaustive search does not work. We have also compared the developed method with the existing connectivity-based methods and FBA-based methods, and show the difference between the solutions of our method and the existing methods. A theoretical analysis of MRI is also conducted, and the NP-completeness of MRI is proved in the Boolean model. Our developed software is available at "http://sunflower.kuicr.kyoto-u.ac.jp/~rogi/minRect/minRect.html."

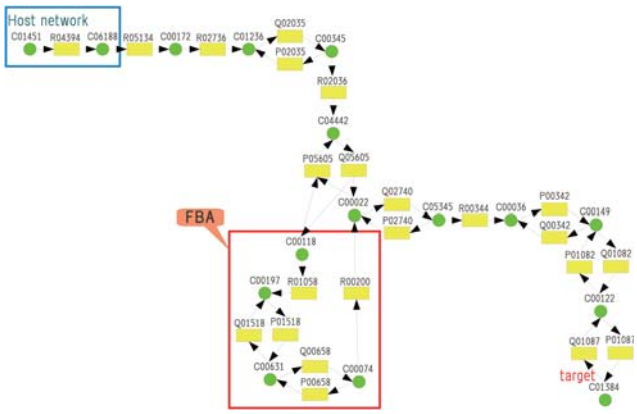


Figure 2. The result of the computer experiment where the host network is E.coli and the target compound is butanol.